DOI:10.16208/j.issn1000-7024.2008.19.033

基于向量相似度计算的半监督的名实体识别

谭红叶[1,2]， 赵铁军[1]， 王浩畅[1]
(1. 哈尔滨工业大学 计算机科学与技术学院，黑龙江 哈尔滨 150001；
2. 山西大学 计算机与信息技术学院，山西 太原 030006)

摘 要：提出一种基于向量相似度计算的半监督的 NER 方法，主要思想是：首先利用 bootstrapping 方法获取 NER 所需的各种特征；然后将待测实例表示为实例特征向量，每一类名实体表示为类特征向量；最后根据每个类特征向量与实例特征向量的相似度进行分类。在人民日报语料上选取疾病名、武器名、交通工具名进行相关测试，F 测度分别为：77.4%，66.1%和 73.1%，结果令人满意。

关键词：名实体识别；特征向量；向量相似度；半监督学习；自举
中图法分类号：TP391.2 **文献标识码**：A **文章编号**：1000-7024 (2008) 19-5047-04

Named entity recognition of semi-superivised based on vector similarity

TAN Hong-ye[1,2], ZHAO Tie-jun[1], WANG Hao-chang[1]
(1. College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
2. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: A semi-supervised method for NER based on vector similarity is presented. The main ideas include: firstly, the features of NE are obtained based on the bootstrapping algorithm, then every instance to be recognized is represented as a feature vector, and so does each category of NE, finally, the vector similarity between the two kinds of vectors is computed, and according to the vector similarity, the final category is decided. The experimental results of disease, weapon and vehicle names on the people daily corpus show that the performances are satisfying, with the F-measure of 77.4%, 66.1% and 73.1% respectively.

Key words: named entity recognition (NER); feature vector; vector similarity; semi-supervised learning; bootstrapping

0 引 言

目前中文名实体的识别(named entity recognition，NER)主要针对人名、地名、机构名 3 大类名实体进行识别，主流方法是采用基于有监督学习的统计模型如：HMM，ME，CRF 等模型。但目前的研究存在一些局限性，主要有：①这 3 类名实体的识别远远不能满足人们对信息自动处理的需求。②采用有监督的学习方法对训练语料过分依赖。而目前，大多中文标注语料只针对人名、地名和机构名进行标注，高质量大规模的扩展名实体的标注语料还不存在。因此如何在未标注语料上对更多类别的名实体进行识别，是本文研究的重点。

无监督学习和半监督学习由于不需要或只需要很少的已标注语料，是克服获取熟语料困难的有效途径。其中，半监督学习是将大规模的未标注语料和少量的已标注语料结合来训练得到一个好的分类器。学者们认为：如果在半监督学习中能够设计出好的模型、特征、相似度函数或目标函数，一定能够克服缺少已标注语料的局限性[1]。

在扩展类别的 NER 中，有研究者采用 bootstrapping 算法进行基于模式的实体识别研究[2-4]。bootstrapping 算法在早期的时候，被研究者归为无监督的学习。但由于 bootstrapping 算法借助了很少的“指导”，即一些“种子”，来进行自学习或自训练，因此近年来被归类为半监督学习[1]。基于 bootstrapping 的 NER 首先利用一些“种子”获取相应类别名实体的一些上下文信息——模式，然后利用这些模式识别新的实例。不断重复这个过程，直到大部分的实例可以被识别。这种方法的实质是基于模式的识别，但由于使用的模式形式单一，不能利用更多的特征，而且模式的匹配只是词语的字面匹配，并要求精确匹配，不能适应自然语言灵活多变的特点，所以取得的性能比较有限，召回率和准确率比较低，以中文人名、地名、机构名的标注为例，召回率分别为 51.6%，79.2%，35.8%，精确率分别为

收稿日期：2007-10-05 E-mail：hytan@mtlab.hit.edu.cn
基金项目：国家自然科学基金项目 (60575041、60473139、60775041)；国家 863 高技术研究发展计划基金项目 (2006AA01Z150)；山西省青年科技基金项目 (20051018)。
作者简介：谭红叶 (1971－)，女，广西灵山人，博士研究生，副教授，研究方向为自然语言处理、信息抽取； 赵铁军 (1962－)，男，黑龙江哈尔滨人，教授，博士生导师，研究方向为自然语言处理、机器翻译； 王浩畅 (1974－)，女，黑龙江哈尔滨人，博士研究生，研究方向为自然信息处理、人工智能。

64.4%，62.0%，40.8%[4]。

基于上述观察，本文提出一种基于向量相似度计算的半监督的 NER 方法，主要思想是：首先利用 bootstrapping 算法获取识别名实体所需要的各种特征；然后把每一个要分类的实例表示为特征向量(称为**实例特征向量**)，每一类名实体也表示为特征向量(称为**类特征向量**)；最后在特征向量空间中，计算向量之间的相似度，选择具有最大相似度的类特征向量对应的类别作为最后的类别。在人民日报语料上选取疾病名、武器名、交通工具名进行相关测试，F 测度分别为：77.4%，66.1%和 73.1%，结果较令人满意。

本文将 NER 转换为特征向量空间中的向量相似度计算问题，主要优点有：①特征向量所包含的特征项可以根据问题进行选择，便于利用更多的特征。②特征向量之间进行相似度计算，避免了模式的严格匹配，适应自然语言灵活多变的特点。③特征向量的获取是在半监督学习方法 bootstrapping 算法的基础上进行，在一定程度上克服了获取大规模标注语料的困难。

1 特征向量的构造

本文将待分类实例和每一类名实体都表示成特征向量。特征向量主要包含两类特征：上下文特征和内部构成特征。具体使用的特征如表 1 所示。其中，所使用的内部构成特征非常简单，只有名实体的尾字或尾词特征。主要原因有：①很多名实体都有一个中心词，这个中心词往往是名实体的尾词。如："哈尔滨工业大学"中的"大学"，"结核病"的"病"，"勃朗宁手枪"中的"手枪"等。而这些尾词恰恰表明了该名实体所属的类别。②名实体的类别很多，每种类别的构成特点不尽相同，用词和用字也很分散。所以不便于使用更多的其它特征。选择特征的时候，为了防止一些"噪音"信息，只有排在前 n 个的，而且出现次数大于 2 次的特征才被选中。

本文的特征向量有两种：实例特征向量$\vec{x}_i$和类特征向量$\vec{c}_j$，实例特征向量是对待分类实例的向量表示，类特征向量是对每一个实体类的向量表示。这两类特征向量都基于上述特征。所有的特征向量构成了特征向量空间V^n。形式化描述如下：

实例特征向量$\vec{x}_i = (w_{i1}, w_{i2}, \cdots, w_{ik}, \cdots, w_{ik})^{\mathrm{T}}$，$w_{ik}$为$\vec{x}_i$中第$k$个特征项的权值。

类特征向量$\vec{c}_j = (w_{j1}, w_{j2}, \cdots, w_{jk}, \cdots, w_{jk})^{\mathrm{T}}$，$w_{jk}$为$\vec{c}_j$中第$k$个特征项的权值。

特征向量空间$V^n = (\vec{x}_1, \vec{x}_2, \cdots \vec{x}_s \vec{c}_1 \vec{c}_2, \cdots \vec{c}_m)$。

实例特征向量的构造可根据待测实例的上下文中提取相应特征，并计算每一个特征的归一化频次作为特征权重来得到。

类特征向量的构造是利用 bootstrapping 算法得到的模式集合来进行。最著名的 bootstrapping 算法主要有 Yarowsky 算法和 Co_Training 算法。其中 Yarowsky 算法是利用一个分类器不断进行自我训练，而 Co_Training 算法是从数据的不同视角 bootstrapping 两个初始分类器，然后在迭代过程中，彼此交换分类结果相互指导学习。我们利用 Yarowsky 算法的思想，在未标注语料上获取实体识别模式。主要步骤为：

(1)建立初始种子集。

(2)根据种子集，在训练语料中抽取窗口大小为 L 的上下文模式，建立候选模式集。

(3)利用一定的评价函数评价和选择模式，生成可用模式集。对于候选模式集中的模式，计算每个模式的分数，并按照分数对模式排序。满足一定阈值条件的模式加入可用模式集中。

(4)利用模式匹配识别相关实体类，构成候选实体名集合。

(5)如果候选实体名集合稳定即不再有新的实体名被识别，或满足一定的迭代次数，则循环终止；否则候选实体名集合将作为新的种子集合，返回步骤(2)开始继续循环。

本文这里与 Yarowsky 算法的主要不同是模式的评价不同。Yarowsky 算法利用概率仅针对模式进行评价。而我们认为bootstrapping过程中，对模式和新识别样例的评价都非常重要。因为利用不可靠的模式识别出的错误实例会被作为新的种子再次进行模式的抽取，这样所抽取的模式必定是错误模式。而模式错误又会被不断传递放大。因此需要在每一次迭代过程中对模式和识别出的实例进行评价，只有满足一定分值的实体名和模式才能进行循环。这种思想与文献[4]的思想类似。本文采用式(1)和(2)进行模式和识别实例的评价。

$$Score(P_j) = 1 - \frac{N_{CommWord}(P_j)}{N_{Word}(P_j)} \tag{1}$$

式中：$N_{CommWord}(P_j)$——模式P_j抽取的普通词个数，$N_{Word}(P_j)$——模式P_j抽取的总词数。

$$Score(NE_j) = \frac{1}{n}\sum_{i=1}^{n} Score(P_i) \tag{2}$$

式中：P_i——本次迭代中可以抽取出实体NE_j的模式，n——可以抽取出实体NE_j的模式的个数。

按照上述思想抽取出的最终可用模式集合包含了实体的各种特征。从该模式集合中抽取表 1 中的特征项，并记录每一个特征项的出现次数，然后计算得到归一化频数，并按照该值排序。选择出现两次以上的特征项作为类特征向量中的特征项，对应的归一化频数作为该特征项的权重。

表 1 系统使用的特征

特征	特征描述
Target_LUnigram	在上文中，与目标词紧邻的一个词
Target_RUniGram	在下文中，与目标词紧邻的一个词
Target_LBigram	在上文中，与目标词紧邻的前两个词
Target_RBigram	在下文中，与目标词紧邻的前两个词
LUniGram	目标词的上文中，所出现的每一个词，不包含 Target_LUnigram
RUniGram	目标词的下文中，所出现的每一个词，不包含 Target_RUnigram
LBigram	目标词的上文中，所出现的每一个 Bigram，不包含 Target_LBigram
RBigram	目标词的下文中，所出现的每一个 Bigram，不包含 Target_RBigram
Central_tail_word	目标词的中心尾词

2 相似度计算和分类判别准则

将待测实例和每一类实体表示成特征向量之后，名实体的识别就被映射成了两个向量的相似度（或距离）计算问题。

计算两个向量$\vec{x}$，$\vec{y}$的相似度$Sim(\vec{x}, \vec{y})$的方法很多，本文采用向量夹角的余弦值来计算相似度，具体计算公式如下

$$Sim(\vec{x}, \vec{y}) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$

根据上述的思想,NER 问题可以形式化描述为:

对于给定数据$\vec{x}$,给定的实体类别集合$C=\{c_i|i=1\cdots n\}$,以及相应的实体类特征向量集合$C'=\{\vec{c_i}|i=1\cdots n\}$。

目标:对数据进行分类。

具体采用的分类判别准则为:

If $Sim(\vec{x},\vec{c_i})>Sim(\vec{x},\vec{c_j})$ and $Sim(\vec{x},\vec{c_i})>\theta$, Then $\vec{x}\in C_i$

其中,$Sim(\vec{x},\vec{c_i})$是根据式(3)计算得到的向量$\vec{x}$和$\vec{c_i}$的相似度,θ为预先设定的阈值。

3 实验结果与分析

手工标注了 1998 年 1 月份和 2 月份的人民日报语料中的疾病名、武器名和交通工具名,并抽取了包含相应的实体名的段落形成测试集,大小是 472k,含 62 184 个词。语料中包含疾病名(Disease)377 个,武器名(Weapon)306 个,交通工具名(Vehilce)251 个。每类实体的初始种子集合不超过 20 个,人工从训练语料中随机选择。在类特征向量的学习过程中采用 1998 年 3~6 月份的语料作为训练语料。训练和识别过程中保留了人民日报语料的分词、人名、地名和机构名的结果,在此基础上进行字母、数字、时间、人名称谓的识别、人名姓与名、复杂地名和复杂机构名的合并。没有利用人民日报语料的词性标注结果,而是在训练过程和识别过程中加入了停用词表,主要包含介词、连词、语气词、助词等虚词,以及一些普通动词。还加入了一些首尾限制,如:首尾不能是时间、字母、数字等。

采用的评价指标是召回率(Recall)、精确率(Precision)和 F-测度(F-measure)。严格地说只有实体边界和类别都正确,才认为识别结果正确。但实验中有些实体类别识别正确,实体边界尽管和答案不一致,但也可以接受。对于这种情形也给予一定的折扣分。这种作法类似于文献[5]中所说的软评测方法。

本文对所提出的方法进行了多组实验:①与其它方法的比较实验;②不同特征对系统性能的影响;③不同上下文窗口大小对系统的影响。实验中决策时用到的阈值$\theta=0.01$。下面将给出实验结果并进行讨论和分析。

(1)与其它方法的比较

为了评价本文所提出方法的有效性,按照文献[4]的方法实现了一个基于 Bootstrapping 的模式匹配的 NER 系统。识别时,区分了上文模式和下文模式,只要与上文模式或下文模式中的一个匹配上就作为识别结果;同时还对模式进行了一定的泛化,如用人名、地名和机构名的类别标记代替原来具体的实体名。在该系统中,实体识别是逐类进行的。两个系统的对比实验结果如表 2 所示,其中 WinLen 指上下文窗口长度。表 2 表明:本文所提出的基于特征向量的方法,即使在没有使用内部特征的情况下也明显好于基于 bootstrapping 的模式匹配的方法。主要原因是基于 bootstrapping 的模式匹配的方法得到的系统只是对模式进行简单的精确匹配,而且所用模式形式比较单一,不能适应自然语言灵活多变的特点。相比之下,基于特征向量的方法不仅可以灵活组合各个特征,而且便于使用更多形式的特征,如:上下文特征、实体内部构成特征等;其次,特征向量之间进行相似度计算,避免了模式的严格匹配,因此保证了识别的准确率和召回率。

(2)不同特征对系统的影响

在 WinLen=3 的情况下测试了各类特征对系统的影响。具体实验结果如表 3 所示。观察表 3,可以看出逐步添加特征的过程中,召回率稳步上升。其次实体的内部尾词特征对系统性能的提升大有帮助,但特征? Bigram 和特征? Unigram 的加入会降低精确率,提升召回率。

(3)不同上下文窗口大小对系统的影响。

上下文窗口长度对名实体的识别有很重要的影响。窗口过小会影响系统的召回率,窗口过大又会影响系统的精确率和效率。表 4 列出了不同上下文长度下的基于向量相似度计算的系统的性能。其中 WinLen 指上下文窗口长度。

从表 4 可以看出,随着上下文窗口逐渐变大,系统的召回率有所提高,但精确率不断下降。系统在 WinLen=1 的时候,各类实体名识别的精确率和总体 F-measure 值最高。但召回率是在窗口大一些的时候,表现较好。应用时应该根据需求调整。

4 结束语

本文提出了一种基于特征向量相似度计算的半监督的实体名识别方法,该方法利用 Bootstrapping 算法得到的上下文模

表 2 不同系统的对比实验结果表 (WinLen = 1)

实体	基于 bootstrapping 的模式匹配的方法			基于特征向量的方法(未使用内部特征)			基于特征向量的方法(使用内部特征)		
	Prec(%)	Rec(%)	$F_{\beta=1}$(%)	Prec(%)	Rec(%)	$F_{\beta=1}$(%)	Prec(%)	Rec(%)	$F_{\beta=1}$(%)
Disease	16.4	35.3	22.4	55.9	40.3	46.8	73.5	81.7	77.4
Weapon	6.4	22.5	10	28.4	46.8	30	57.5	77.8	66.1
Vehicle	3.6	35.9	6.5	30	16.7	21.5	62.1	88.8	73.1

表 3 利用不同特征识别实体的结果 (WinLen = 3)

特征	Disease		Weapon		Vehicle	
	Prec(%)	Rec(%)	Prec(%)	Rec(%)	Prec(%)	Rec(%)
T? Unigram	60.1	41.2	42.7	32.7	32.5	50.1
T? Unigram+T? Bigram	61.5	47.3	41.4	34.4	33.1	53.8
T? Unigram+T? Bigram+? Bigram+? Unigram	51.4	61.7	38.3	51.7	27.3	68.4
T? Unigram+T? Bigram+? Bigram+? Unigram+Tail Word	62.5	84.9	48.3	78.8	34.5	89.1

解释:T? Unigram 指 TLUnigram 和 TRUnigram; T? Bigram 指 TLBigram 和 TRBigram;
? Bigram 指 LBigram 和 RBigram; ? Unigram 指 LUnigram 和 RUnigram;

表 4　不同上下文窗口长度下实体识别结果

WinLen	Disease(%)			Weapon(%)			Vehicle(%)		
	Prec	Rec	F_val	Prec	Rec	F_val	Prec	Rec	F_val
1	73.5	81.7	77.4	57.5	77.8	66.1	62.1	88.8	73.1
2	73.0	83.3	77.8	50.1	79.1	61.3	44.2	88.9	59.1
3	62.5	84.9	72.0	48.3	78.8	59.9	34.5	89.1	49.7
4	63.4	85.4	72.8	40.5	80.7	53.9	31.8	89.3	46.8
5	59.0	85.9	70.0	37.0	80.1	50.6	29.9	89.5	44.8

式集合获取实体类特征向量，把待测实例表示成实例特征向量，然后计算待测实例特征向量与每个类特征向量之间的相似度，选择具有最大相似度的类特征向量的所属类别作为最后的实体类别。该方法由于有效地利用了各种特征，避免了模式的严格匹配，取得了令人满意的效果。而且还实现了利用未标注语料进行名实体的识别，克服了获取大规模标注语料的困难。目前系统性能距离实用还有很大距离，仍然有大量细致的工作要做：①在规模更大的语料上进行相关测试。②尝试一些后处理技术，提高系统的性能。

参考文献:

[1] Zhu X J. Semi-supervised learning literature survey [EB/OL]. http://www.cs.wise.edu/~jerryzhu/pub/ssl_survey.pdf.

[2] Yangarber R,Lin W,Grishman R.Unsupervised learning of generalized names[C].Taipei, Taiwan:Proceedings of the 19th International Conference on Computational Linguistics,2002:1135-1141.

[3] Lin W, Yangarber R, Grishman R. Bootstrapped learning of semantic classes from positive and negative examples[C].Washington DC,USA:Proceedings of ICML Workshop on the Continuum from Labeled to Unlabeled Data,2003:103-101.

[4] Thelen M, Riloff E.A bootstrapping method for learning semantic lexicons using extraction pattern contexts[C]. Philadelphia, USA:Proceedings of the Conference on Empirical Methods in Natural Language Processing,2002:214-221.

[5] 刘非凡,赵军.面向商务信息抽取的产品命名实体识别研究[C].自然语言理解与大规模内容计算.北京:清华大学出版社,2005:415-421.

[6] Satoshi Sekine,Named Entity.History and future[EB/OL]. http://cs.nyu.edu/sekine/papers/.

[7] 张华平,刘群.基于角色标注的中国人名自动识别研究[J].计算机学报,2004,27(1):85-91.

[8] 赵健.条件概率模型及其在中文名实体识别中的应用研究[D].哈尔滨:哈尔滨工业大学计算机学院,2006.

(上接第 5015 页)

诊情况(a_1)、X 胸片情况(a_3)、痰培养情况(a_5)这 3 个因素就足够了。而利用文献[2]算法则需要考虑每个对象的肺部听诊情况(a_1)、体温情况(a_2)、X 胸片情况(a_3)、痰培养情况(a_5)这 4 个因素。由此可以看出在计算时间相同的情况下，利用本文算法比一般启发式算法更加简化了分类规则的提取，提高决策效率。这在实际应用中具有很高的价值。

对于一些比较特殊的情况，如对于一个给定的决策表，求得 $I(C_0,D) = I(C,D)$时，利用本文算法，此时经判断已满足循环终止条件；故直接输出条件属性集 C 相对于决策属性集 D 的一个相对约简 $B = C_0$。与一般启发式属性约简算法相比，此时利用本文算法不仅能够求出 C 相对于 D 的一个相对约简，同时在求得相对约简中不存在冗余属性(由定理知)。

3　结束语

本文通过对一般启发式属性约简算法进行分析，针对算法中存在的不足，提出了一种基于条件熵的启发式属性约简算法。在算法中加入了消除冗余的二次约简过程，同时在一些特殊的情况下(如：条件属性集 C 相对于决策属性集 D 的核 C_0 为空集与 $I(C_0,D)= I(C,D)$)，本文算法也给出了相应的处理方法。与一般启发式属性约简算法相比，本文算法在任何情况下都能够求得条件属性集 C 相对于决策属性集 D 的相对约简，且在求得的相对约简中不存在冗余属性。最后通过真实数据的实验结果验证了算法的有效性和约简效果。

参考文献:

[1] Wong S K M,Ziarko W.On optional decision rules in decision tables [J]. Bulletion of Polish Academy of Sciences, 1985,33 (11/22):693-696.

[2] 苗夺谦,胡桂荣.知识约简的一种启发式算法[J].计算机研究与发展,1999,36(6):681-684.

[3] 苗夺谦,王珏.粗糙集理论中概念与运算的信息表示[J].软件学报,1999,10(2):113-116.

[4] 胡丹,莫智文.关于粗糙集理论与信息熵的几点注记[J].四川师范大学学报(自然科学版),2002,25(3):257-260.

[5] 常犁云, 王国胤. 一种基于 Rough Set 理论的属性约简及规则提取方法[J].软件学报,1999,10(11):1206-1211.

[6] 胡可云,陆玉昌,石纯一.粗糙集理论及其应用进展[J].清华大学学报(自然科学版),2001,41(1):64-68.

[7] 石峰,娄臻亮.一种改进的粗糙集属性约简启发式算法[J].上海交通大学学报,2002,36(4):478-481.

[8] 瞿彬彬,卢炎生.基于粗糙集的属性约简算法研究[J].华中科技大学学报(自然科学版), 2005,33(8):30-33.

[9] 王卫玲,刘培玉.一种改进的基于条件互信息的特征选择算法[J].计算机应用,2007,27(2):433-435.

[10] 王国胤. Rough 集理论与知识获取[M].西安:西安交通大学出版社,2001:117-145.